MATERIAL EXLUVISO
GRUPO CURSO FÁCIL
CLIQUE AQUI
E ACESSE + MATERIAIS







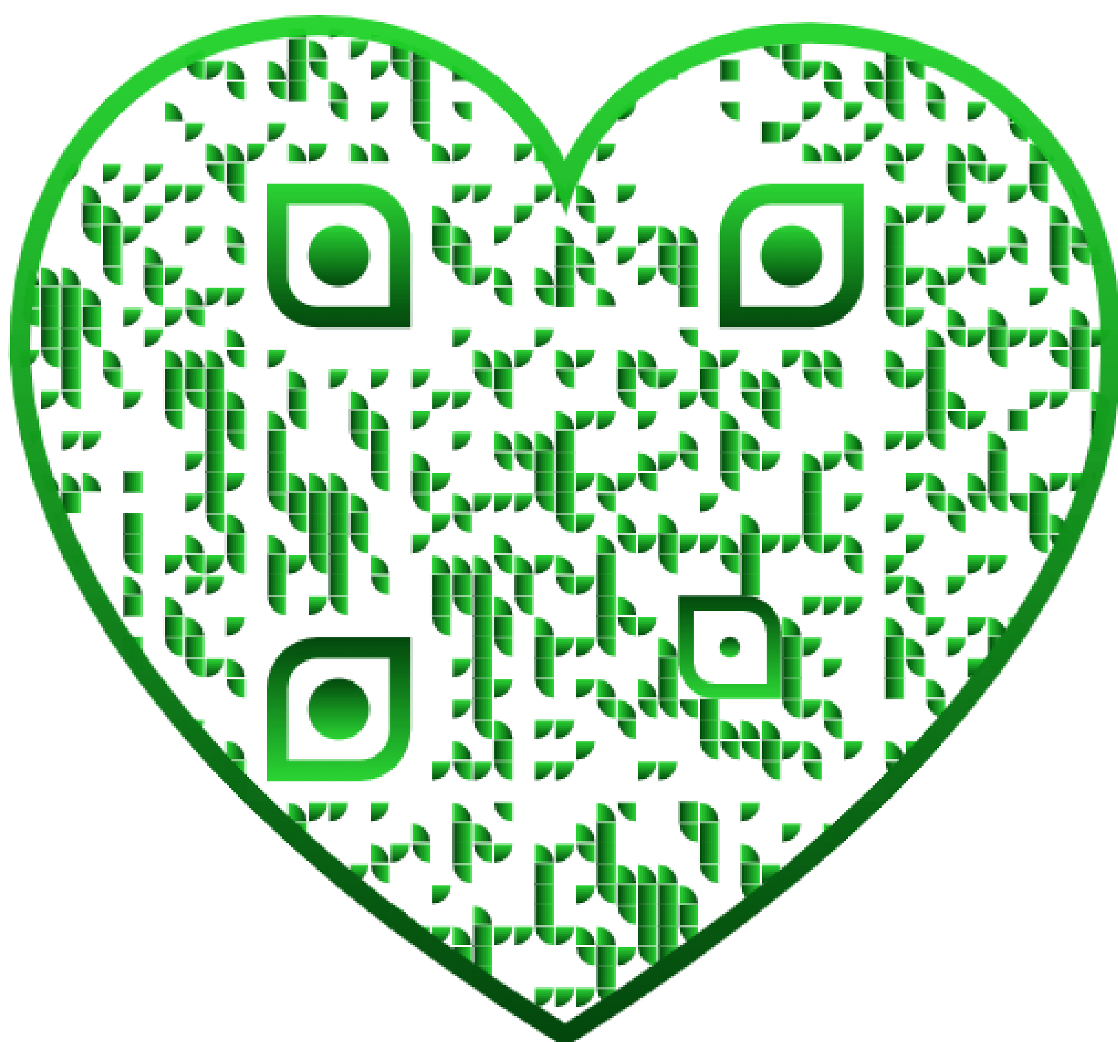
Informática para Caixa Econômica Federal - CEF (Técnico Bancário)

Prof. Diego Carvalho, Prof. Marcos Vinícius Alves Franco, Prof. Diego Carvalho

12 Análise de Informações - BI e Data Warehouse

Documento última vez atualizado em 02/01/2024 às 19:55.





Índice

12.1) Análise de Informações - BI e Data Warehouse - Sistemas de Suporte à Decisão 3

12.2) Análise de Informações - BI e Data Warehouse - Data Warehouse (DW) 16

12.3) Análise de Informações - BI e Data Warehouse - Resumo 36

12.4) Mapa Mental - Análise de Informações - BI e Data Warehouse 41

12.5) Questões Comentadas - Análise de Informações - BI e Data Warehouse - Multibancas 42

12.1 Análise de Informações - BI e Data Warehouse - Sistemas de Suporte à Decisão

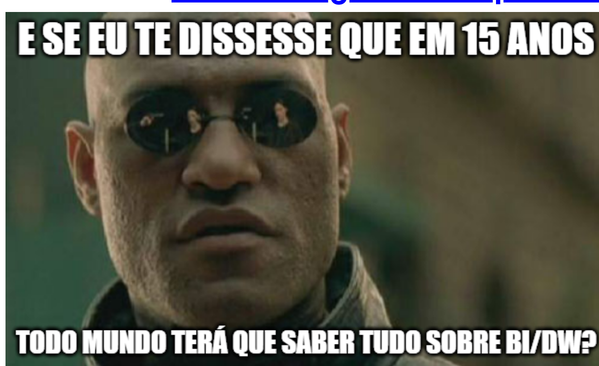


Para assistir ao vídeo correspondente, acesse o LDI.

Apresentação da Aula

Fala, galera! O assunto da nossa aula de hoje **Business Intelligence e Data Warehouse**. Essa aula é um pouco pesada, mas é necessário estudá-la porque ela é a base para a Mineração de Dados. Aqui veremos um novo tipo de banco de dados, como ele é construído e porque ele é utilizado. Eu juro que tentei ao máximo deixar o mais palatável possível para vocês, mas essa é uma aula que exigirá bastante atenção! Levantem, tomem um energético e venham comigo...

PROFESSOR DIEGO CARVALHO - www.instagram.com/professordiegocarvalho [1]



Galera, todos os tópicos da aula possuem Faixas de Incidência, que indicam se o assunto cai muito ou pouco em prova. *Diego, se cai pouco para que colocar em aula?* Cair pouco não significa que não cairá justamente na sua prova! A ideia aqui é: se você está com pouco tempo e precisa ver somente aquilo que cai mais, você pode filtrar pelas incidências média, alta e altíssima; se você tem tempo sobrando e quer ver tudo, vejam também as incidências baixas e baixíssimas. *Fechado?*

INCIDÊNCIA EM PROVA: baixíssima
INCIDÊNCIA EM PROVA: baixa
INCIDÊNCIA EM PROVA: média
INCIDÊNCIA EM PROVA: ALTA
INCIDÊNCIA EM PROVA: Altíssima

Além disso, essas faixas não são por banca – é baseado tanto na quantidade de vezes que caiu em prova independentemente da banca e também em minhas avaliações sobre cada assunto...

#ATENÇÃO

Avisos Importantes

O curso abrange todos os níveis de conhecimento...

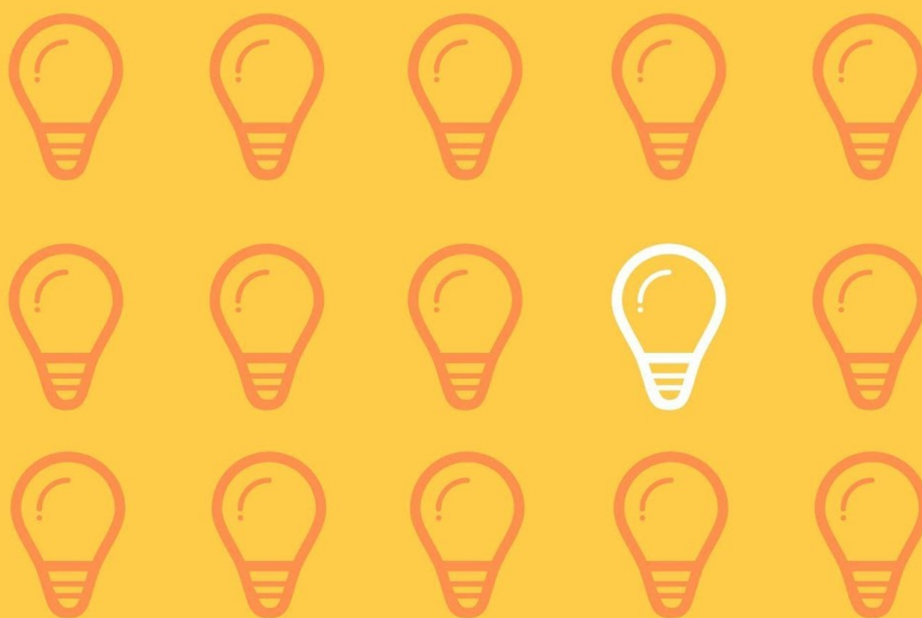
Esse curso foi desenvolvido para ser acessível a **alunos com diversos níveis de conhecimento diferentes.** Temos alunos mais avançados que têm conhecimento prévio ou têm facilidade com o assunto. Por outro lado, temos alunos iniciantes, que nunca tiveram contato com a matéria ou até mesmo que têm trauma dessa disciplina. A ideia aqui é tentar atingir ambos os públicos - iniciantes e avançados - da melhor maneira possível..

Por que estou enfatizando isso?

O **material completo** é composto de muitas histórias pessoais, exemplos, metáforas, piadas, memes, questões, desafios, esquemas, diagramas, imagens, entre outros. Já o **material simplificado** possui exatamente o mesmo núcleo do material completo, mas ele é menor e mais objetivo. *Professor, eu devo estudar por qual material?* Se você quiser se aprofundar nos assuntos ou tem dificuldade com a matéria, necessitando de um material mais passo-a-passo, utilize o material completo. Se você não quer se aprofundar nos assuntos ou tem facilidade com a matéria, necessitando de um material mais direto ao ponto, utilize o material simplificado.

Por fim...

O curso contém diversas questões espalhadas em meio à teoria. Essas questões possuem um comentário mais simplificado porque **têm o único objetivo de apresentar ao aluno como bancas de concurso cobram o assunto previamente administrado**. A imensa maioria das questões para que o aluno avalie seus conhecimentos sobre a matéria estão dispostas ao final da aula na lista de exercícios e **possuem comentários bem mais abrangentes**.



• ATENÇÃO •

Existem muitos exercícios sobre esse tema em sites de questões, no entanto a imensa maioria foi aplicada em provas para cargos específicos de Tecnologia da Informação (TI), os quais podem demandar um conhecimento muito mais aprofundado da matéria.

Dessa forma, recomendo que vocês tenham muita atenção na seleção das questões realizadas para que não extrapolem o nível cobrado na sua prova.

Qualquer dúvida, estou à disposição para maiores esclarecimentos!

Sistemas de Suporte à Decisão

Conceitos Básicos

INCIDÊNCIA EM PROVA: baixa

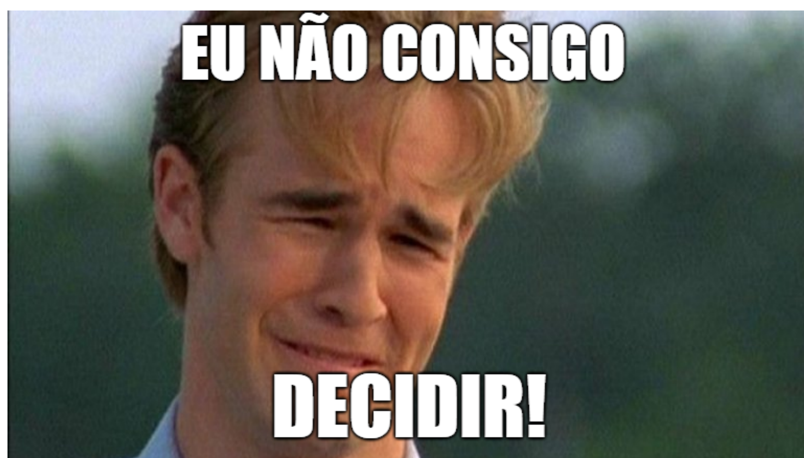


Vamos, galera... gás totaaaaaal! Hoje eu estou animado, vamos matar mais um assunto do seu edital! **Por falar nisso, o assunto da aula de hoje são os Sistemas de Suporte à Decisão**! Galera, decidir sobre algo nem sempre é fácil! Eu vou dar um exemplo: é terça-feira à noite, minha noiva diz que está morrendo de fome. Segue o diálogo apresentado ao lado...

Quem nunca ficou horas vendo as opções de restaurante no iFood? Quem nunca viu todos os cartazes de filme da Netflix tentando escolher algum para assistir com a namorada/namorado e acabou não assistindo nada? Pois é, quando se tem muitas opções, **a escolha acaba se tornando mais difícil**. Se isso vale para escolhas simples, imagine para grandes decisões!

Pessoal, essas decisões que eu mencionei ocorrem várias vezes durante nossa vida e não têm nenhuma repercussão mais grave. Assistir um filme chato não vai me fazer parar de ver filmes no futuro. Comer algo que eu não quero também não me causará nenhum dano (e ainda ajuda nos relacionamentos). No entanto, **há algumas decisões que ocorrem basicamente uma ou, no máximo, duas vezes na vida e geralmente podem causar grandes impactos em sua vida**.

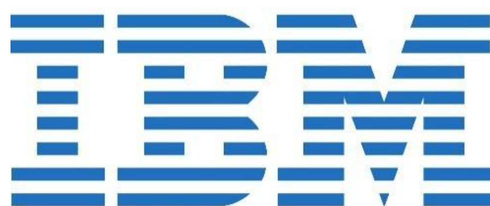
Dá um exemplo, professor? **Um exemplo clássico é sua carreira! Quando você tem apenas dezessete anos, você tem que decidir qual será sua carreira** (por vezes, isso valerá para o resto da vida). Quando eu fiz vestibular, eu fiquei dias e dias pensando em qual curso escolher: Ciência da Computação ou Economia. *Eu poderia desistir no meio do curso e fazer um novo vestibular?* Sim, mas – em muitos casos – isso pode não ser viável. *Concordam?*



Outro exemplo: eu sou um cara meio enrolado! Eu precisei namorar oito anos para finalmente tomar coragem e pedir minha namorada em casamento. **Essa é uma decisão que se faz uma ou duas vezes na vida** – exceto se você for a Gretchen ou o Vinicius de Moraes! *E se vocês passarem para um concurso em um local muito distante de onde vocês moram? Vão se mudar e deixar amigos e parentes? E se dois concursos importantes terão provas no mesmo dia?*

Uma particularidade de todos esses exemplos é que eles trataram de decisões individuais! *No entanto, o que aconteceria se estivéssemos falando de uma organização – seja ela pública ou privada?* **Pois é, decisões individuais afetam apenas o indivíduo, mas decisões organizacionais podem afetar centenas, milhões ou até bilhões de pessoas**. *Vocês querem saber de um ótimo exemplo?* A assinatura ilimitada!

Vocês já pararam para pensar quanto os sócios-fundadores do Estratégia Concursos devem ter pensado antes de decidirem migrar para um plano de negócio baseado em assinatura ilimitada? **Pois é, esse é um tipo de decisão importantíssima para uma organização que pode levá-la a outro patamar ou pode levá-la à ruína**. Aliás, o mundo está repleto de exemplos de como decisões erradas podem levar a prejuízos colossais. Vejam só...



- **Decisão:**

Comprar um sistema operacional da Microsoft para seus computadores pessoais.

- **Erro:**

A IBM não previu que se tornaria refém da Microsoft.

- **Prejuízo:**

US$ 75.000.000.000,00



- **Decisão:**

Aquisição da concorrente Compaq

- **Erro:**

A união das empresas não funcionou e os lucros permaneceram inalterados.

- **Prejuízo:**

US$ 19.000.000.000,00



- **Decisão:**

Investir em linhas internacionais, como a rota São Paulo a Frankfurt (Alemanha).

- **Erro:**

Superestimou a demanda dos usuários.

- **Prejuízo:**

R$ 40.000.000,00

Pessoal, para evitar esses prejuízos surreais e ajudar a tomar melhores decisões, foram criados os Sistemas de Suporte/Apoio à Decisão. Eles tratam basicamente de sistemas que suportam o processo de tomada de decisão de organizações. *Professor, ele toma decisões sozinho?* Não, ele apoia, auxilia, guia um gestor – mas não substitui a intervenção humana! Na verdade, você utiliza ferramentas ou recursos que te auxiliam na tomada de decisões o tempo todo. *Quer ver?*

Quando você precisa decidir o que comer, você olha no relógio para verificar as horas! Se for de manhã, você provavelmente não vai comer uma feijoada. Se for no almoço, você provavelmente não vai comer apenas uma sopa. Você também verifica sua carteira ou seu saldo bancário para verificar se você tem dinheiro suficiente ou não. **Se você tem apenas cinco reais, o máximo que dá para comer é um churrasquinho na esquina ou um cachorro quente**.

Quando eu vou comer, eu olho sempre para minha barriga. Se ela estiver imensa, eu decido comer uma saladinha leve; mas se estiver tranquila, eu mando aquela rabada com agrião sem peso na consciência. **O que eu quero mostrar é que nós já utilizamos ferramentas ou recursos para nos auxiliar na tomada de decisões**. Já organizações precisam evidentemente de ferramentas mais parrudas. *Por que?* Porque em geral decisões necessitam de cálculos, muitos cálculos!

Quando você vê em seu smartphone a previsão do tempo para amanhã, por trás dos panos há um supercomputador com altíssima capacidade de processamento realizando trilhões de cálculos para chegar em um resultado. **Ele analisa umidade, pressão, histórico e centenas de outras variáveis para alcançar uma previsão mais acurada**. Decisões em geral também necessitam de muitos dados – geralmente extraídos de um Data Warehouse (veremos à frente).

Querem ver um exemplo? Em 2015, a Apple lançou seu Apple Watch! Vocês sabem que há pessoas que são muito fãs de produtos da Apple e fazem filas enormes para comprá-los – **algumas vezes sem sequer saber como eles serão**. Por outro lado, a Apple também não sabe exatamente qual será a demanda pelos seus novos produtos antes de lançá-los, logo é complicado estimar quantos devem ser produzidos de antemão para que as lojas tenham estoque para todos os compradores.

Vocês já imaginaram pagar caro por um Apple Watch e só recebê-lo após dois meses? **Isso seria péssimo para a imagem da empresa**. Além disso, deve-se levar em consideração que o lançamento ocorre por vezes em escala mundial, logo a empresa tem que estimar quantos devem ser produzidos considerando onde o produto será lançado, o poder de compra da comunidade local, o histórico de compras de outros produtos da empresa, entre outros.

Em suma, a empresa tem que decidir quantos serão fabricados antes do lançamento e qual a proporção de distribuição em cada país. Exemplo: 40% do estoque ficará nos EUA porque o poder de compra dos americanos é alto e o histórico é favorável; somente 5% será enviado ao Brasil porque o poder de compra dos brasileiros é baixo e as pessoas em geral não têm o costume de investir esse valor em um relógio. Para tomar essas decisões, são necessários muitos dados!

Ainda não acabou! Decisões necessitam de alta consistência. *Como assim, Diego?* **Dados os mesmos conjuntos de informações de entrada, um sistema deve obter resultados semelhantes**. Vamos imaginar – por exemplo – que você está precisando de uma grana e decide ir a um banco pedir um empréstimo. *Vocês já imaginaram se a quantidade de crédito oferecida dependesse por exemplo do humor do gerente naquele dia específico?* Seria complicado...

Sistemas são mais consistentes que humanos, portanto nós podemos programá-los para seguir diversas regras e sempre aplicar os mesmos critérios para atingir resultados consistentes. Logo, se dez pessoas com remunerações parecidas, quantidade de poupança semelhantes, idades iguais e nenhum histórico de dívidas forem a um banco pedir um empréstimo, o sistema de suporte a decisão deve sugerir valores de crédito semelhantes para as dez pessoas.

Por fim, decisões frequentemente devem ser tempestivas! Em outras palavras, quando você precisa tomar uma decisão, você tem que tomá-la rapidamente. Estou eu aqui pensando em um bom exemplo para vocês quando vejo uma notícia: **Capitania dos Portos alerta para possível ciclone marítimo com ventos de até 89 km/h no sul e baixo sul da Bahia**. *Você é o governador da Bahia agora, o que você faria?*



Capitania dos Portos alerta para possível ciclone marítimo com ventos de até 89 km/h no sul e baixo sul da Bahia

Fenômeno deve atingir cidades como Itamaraju e Valença. Outras regiões do estado têm previsão de chuvas e rajadas de ventos, mas não é em decorrência do ciclone.

Por Thaic Carvalho, G1 BA
22/03/2019 14h42 · Atualizado há 3 dias



Sistemas de Suporte à Decisão podem auxiliar a decidir qual meio deve ser utilizado para alertar os moradores (mensagens, ligações telefônicas, avisos sonoros?); podem auxiliar a decidir para onde mandar ajuda (para cidades onde efetivamente passará o ciclone ou para cidades ao lado que possuam maior infraestrutura?); qual ajuda mandar (comida, água,

remédios?). **Se você demorar alguns minutos, pessoas podem morrer! Logo, decisões devem ser rápidas!**

Enfim, galera... **as decisões que passam por esse tipo de sistema geralmente são resultado de muitos cálculos, necessitam de um volume altíssimo de dados, são altamente consistentes e devem ser tempestivas**. *Legal?* Bem... feita essa introdução, é importante ver uma definição formal: trata-se de um conjunto de procedimentos baseados em modelos para processamento de dados e julgamentos para ajudar o seu utilizador a tomar decisões.

Business Intelligence (BI)

INCIDÊNCIA EM PROVA: ALTA



Os tempos foram passando, o mundo foi mudando e o termo Sistema de Suporte à Decisão se tornou o que chamamos hoje de Business Intelligence (BI)! O princípio é o mesmo, porém com alguns detalhes que veremos ao longo da aula. **Vamos ver na tabela a seguir diversas definições formais bastante comuns em prova – percebam como há algumas palavras-chave em comum em quase todas essas definições.**

DEFINIÇÕES
Business Intelligence é o processo de recolhimento e tratamento de informações que apoiarão a gestão de um negócio.
Palavras-chave: processo; tratamento de informações; apoio à gestão de negócio.
Business Intelligence é o processo inteligente de coleta, organização, análise, compartilhamento e monitoração de dados que, depois de processados, geram informações para o suporte e para a tomada de decisões no ambiente de negócios.
Palavras-chave: processo inteligente; coleta, organização, análise, compartilhamento e monitoração de dados; tomada de decisões.
Business Intelligence trata da capacidade de cruzar informações de diferentes bancos de dados, gerando relatórios analíticos diversos.
Palavras-chave: cruzamento de informações; diferentes bancos de dados; relatórios analíticos.
Business Intelligence se refere à aplicação de técnicas analíticas para informações sobre condições de negócio no sentido de melhorá-las, de uma maneira automatizada, mas com a interpretação e respostas humanas, de forma a melhorar a tomada de decisões.
Palavras-chave: técnica analíticas; automatizada; tomada de decisões;
Business Intelligence trata da descrição de habilidades das organizações para acessar dados e explorar informações, analisando e desenvolvendo percepções e entendimentos a seu respeito.
Palavras-chave: habilidade; explorar informações; análise de percepções e entendimentos.
Business Intelligence são técnicas, métodos e ferramentas de análise de dados para subsidiar processos de decisão de uma empresa.
Palavras-chave: técnicos, métodos e ferramentas; análise de dados; processo de decisão.
Business Intelligence nada mais é que uma série de conceitos e metodologias para auxiliar na tomada de decisões estratégicas nas empresas através principalmente de geração de relatórios gerenciais analíticos.

Palavras-chave: conceitos e metodologias; tomada de decisões estratégicas; geração de relatórios analíticos.
Business Inteligence é um termo utilizado para descrever as habilidades das corporações para coletar dados e explorar informações, analisá-las e desenvolver entendimentos para tomada de melhores decisões.
Palavras-chave: habilidades; coleta, exploração e análise de dados; tomada de decisão.

Agora que nós já sabemos tudo isso, nós podemos criar a nossa própria definição:

Business Intelligence (Inteligência de Negócio) é um termo criado pelo Gartner Group utilizado para representar um conjunto de processos, técnicas, metodologias, habilidades, ferramentas e capacidades utilizadas para acessar, coletar, organizar, tratar, analisar, cruzar, processar, compartilhar e monitorar dados de diversas fontes com o intuito de gerar informações e relatórios analíticos que suportem a gestão corporativa, a definição de estratégias e a tomada de decisão em ambientes de negócio.

O Business Intelligence é realizado sobre dados históricos consolidados, provenientes de bases multidimensionais, tais como Data Warehouses. *O que é isso, professor?* Calma, pequeno gafanhoto! Nós veremos isso tudo ainda na aula de hoje. **O que vocês precisam guardar é que o Business Intelligence é capaz de fornecer uma visão dos dados facilitando análises diagnósticas, descritivas e até preditivas para suportar a tomada de decisão** [2]. Vejamos...

- **Análise descritiva:** é o exame de dados ou conteúdos para responder à pergunta: "*O que aconteceu?*" ou "*o que está acontecendo?*". É caracterizado pelo Business Intelligence tradicional e visualizações como gráficos de pizza, gráficos de barras, etc que servem para que gestores tomem decisões;

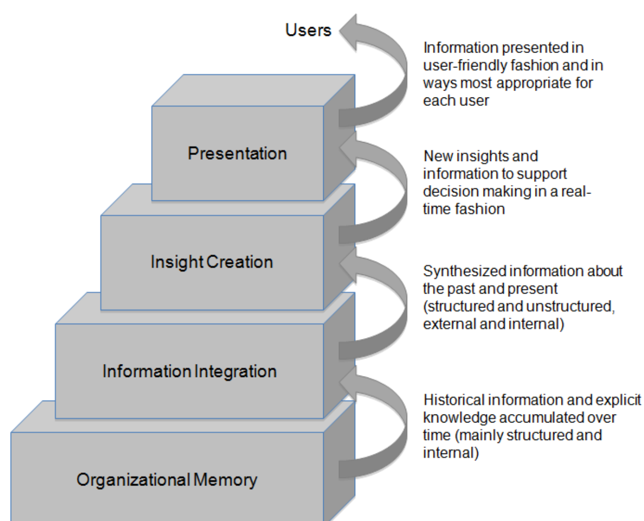
- **Análise Diagnóstica:** é uma forma de análise avançada que examina dados ou conteúdos para responder à pergunta: "*Por que isso aconteceu?*", e é caracterizada por técnicas como detalhamento, descoberta de dados, mineração de dados e correlações. Esse é o tipo de análise empregada em investigações de causa-raiz;

- **Análise Preditiva:** combina técnicas de estatística, mineração de dados e aprendizagem de máquina (*Machine Learning*) para encontrar significado em grandes quantidades de dados, trabalhando com probabilidades, entre outros para antecipar comportamentos futuros com base em eventos passados. Responde à pergunta: "*O que vai acontecer?*";

- **Análise Prescritiva:** é uma forma de análise avançada que examina os dados ou os conteúdos para responder à pergunta: *"O que deve ser feito?"* ou *"O que podemos fazer caso algo aconteça?"*, e é caracterizada por técnicas como análise de gráficos, simulação, redes neurais, mecanismos de recomendação, heurísticas, aprendizado de máquina, etc.

É interessante mencionar que um usuário que desejar utilizar os conceitos de BI precisará de uma infraestrutura arquitetural específica capaz de extrair, limpar, formatar, transformar e carregar dados estruturados ou não estruturados de diversas fontes em depósitos de informações que possam ser acessados por sistemas analíticos. **Além disso, BI possui quatro habilidades: memória organizacional, informação integrada, criação de conhecimento (*insight*) e apresentação.**

O principal recurso de Business Intelligence é a Memória Organizacional. A informação e o conhecimento são armazenados no sistema de BI de forma que possam ser acessados e observados posteriormente. Isso ajuda na próxima habilidade de Business Intelligence: Integração de Informações, que é a habilidade de centralizar informações de diversas fontes. Informações centralizadas em um único local ajudam na Criação de Insights.



Criar insights significa desenvolver a capacidade de ter intuições sobre o negócio para ajudar a melhorar a tomada de decisões. Por fim, a criação de insights fornece contribuições para o recurso de Apresentação. **Por meio dela, o sistema de Business Intelligence tem a habilidade de gerar relatórios e ferramentas adequados, legíveis e inteligíveis para a maioria dos usuários das ferramentas**.

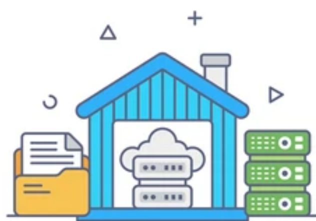
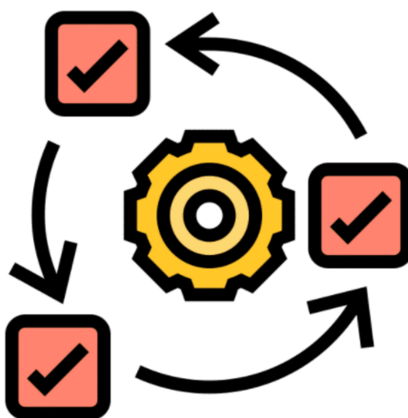
O próprio Gartner Group menciona essas habilidades quando define BI: trata-se das habilidades das corporações de acessar dados e explorar as informações (normalmente contidas em um Data Warehouse ou Data Mart), analisando-as e desenvolvendo percepções e entendimentos a seu respeito – o que as permite incrementar e tornar a tomada de decisões mais pautada em informações. *Entendido?*

Por fim, vamos falar sobre a Arquitetura de Business Intelligence! **De acordo com E. Turban, BI possui quatro grandes componentes:** (1) Data Warehouse, com seus dados-fonte; (2) Análise de Negócio, uma coleção de ferramentas para manipular e analisar os dados do Data Warehouse, incluindo Data Mining; (3) Business Performance Management (BPM), para monitoria e análise do desempenho; e (4) uma interface de usuário, como um *dashboard*.

Observe que o ambiente de Data Warehousing é sobretudo de responsabilidade de uma equipe técnica, e o ambiente de análise (também conhecido como análise de negócios) está no âmbito dos usuários de negócios. Qualquer usuário pode conectar-se ao sistema por meio de uma interface de usuário, como um navegador, e a alta administração pode usar o componente de BPM e também o dashboard. Vejamos uma imagem dessa arquitetura de alto nível...

Componentes	Descrição
DATA WAREHOUSE	Partindo do lado esquerdo da figura, vemos um fluxo de dados dos sistemas operacionais (Ex: CRM, ERP, etc) até um Data Warehouse, que é um banco ou repositório de dados de interesse dos gestores preparado para dar suporte a aplicações de tomada de decisão. As aplicações variam de simples gerações de relatórios ou consultas a complexas otimizações.
ANÁLISE DE NEGÓCIOS	Trata-se de um conjunto de ferramentas para manipular, minerar e analisar dados, criar relatórios, realizar consultas sob demanda, entre outros. Dessa forma, usuários de negócio são capazes de identificar com rapidez e facilidade as tendências de desempenho (Ex: usuários podem isolar e identificar produtos, clientes ou regiões que apresentam tendências significativas de alta ou de baixa, ou que constituem fontes de problemas).
BPM	Este componente permite otimizar o desempenho geral de uma organização por meio da conexão de métricas (Ex: informações financeiras) com desempenhos reais da organização. Para tal, utiliza-se da análise, geração de relatórios e consultas para comparar o desempenho de negócios às metas estabelecidas – além de oferecer uma plataforma para compartilhar metas de desempenho e resultados da empresa.
Interface de usuário	Esse componente permite uma visão rápida dos dados por meio de ferramentas de visualização, como dashboards, painéis, portais, gráficos, cockpits, etc. Os dashboards fornecem uma visão abrangente das medidas, tendências e exceções do desempenho corporativo provenientes de

	múltiplas áreas do negócio, propiciando uma visão imediata da saúde da organização.

ARQUITETURA BI	 	**Trata-se de uma coleção de dados orientados por assunto, integrados, variáveis com o tempo e não-voláteis, para dar suporte ao processo de tomada de decisão.**
		**TRATA-SE DE um conjunto de ferramentas para manipular, minerar e analisar dados, criar relatórios, realizar consultas sob demanda, entre outros.**
		**Trata-se de um conjunto de ferramentas que permitem comparar e compartilhar o desempenho de negócios de uma organização em relação às metas estabelecidas.**
		**Trata-se de um conjunto de ferramentas que permitem a rápida e abrangente visualização de dados por meio de dashboards e softwares similares.**

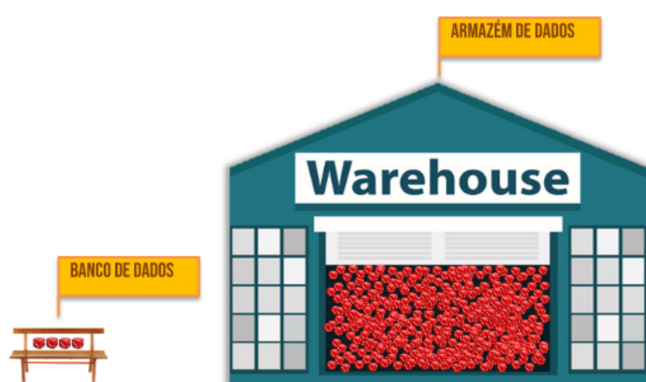
12.2 Análise de Informações - BI e Data Warehouse - Data Warehouse (DW)

 Para assistir ao vídeo correspondente, acesse o LDI.

Data Warehouse (DW)
Contexto Geral

INCIDÊNCIA EM PROVA: MÉDIA

Pessoal, nós vamos iniciar agora um estudo um pouco mais aprofundado. Vocês descobrirão várias tecnologias novas, diversos nomes diferentes e – como o assunto vai se tornar bem mais técnico – eu vou ter que utilizar mais analogias e exemplos para que vocês possam absorver com maior facilidade. **Dito isso, eu vou fazer uma grande contextualização inicial para explicar o assunto da nossa aula para que vocês vejam que não é nenhum bicho de sete cabeças...**

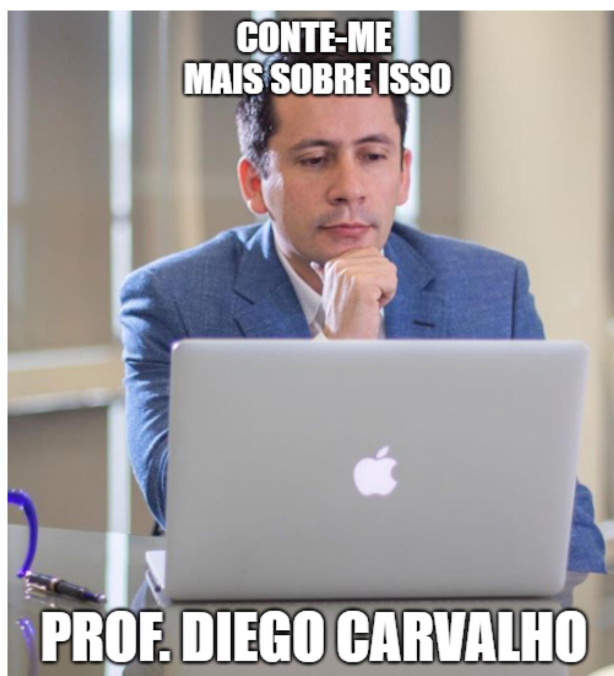


Data Warehouse significa Armazém de Dados! Notem que é algo maior que um banco de dados... é um **armazém** de dados! *Professor, para que eu vou precisar de um banco de dados enorme?*

Vamos imaginar um cenário hipotético: Quatro servidores públicos decidem criar uma empresa de cursos preparatórios para concursos! São poucos funcionários, poucas vendas e todas elas podiam ser registradas em um pequeno banco de dados por meio de um sistema. *E não é que essa empresa começa a dar muito certo?*

Começa, então, a surgir alunos de todo canto do país querendo comprar cursos e mais cursos. *Resultado?* O volume de vendas aumentou substancialmente – assim como os dados armazenados em nosso pequeno banco de dados. Além disso, os sócios começaram a ficar estafados de tanto trabalhar e começaram a contratar vários funcionários para trabalhar na área financeira, comercial, recursos humanos, pedagógica, coaching, etc.

Professor, esse tanto de novos funcionários também utilizavam o mesmo sistema e esse mesmo banco de dados? Sim, senhor! No entanto, os bancos de dados tradicionais possuem uma característica específica: eles são extremamente eficientes para trabalhar com transações.



Vocês se lembram o que é uma transação? **Basicamente é qualquer leitura, inclusão, alteração ou exclusão de dados**. *Como assim, Diego?* Vamos abrir um parêntese aqui!



Galera, quando nós lançamos um curso para vocês, nós apresentamos um cronograma de aulas. Quando chega a data específica, nós inserimos a aula no banco de dados (**inclusão**); vocês entram em nosso site e acessam a aula do banco de dados (**leitura**); se eu acabo inserindo a aula errada, eu posso deletá-la do banco de dados (**exclusão**); ou se há algum erro na aula, eu posso editá-la no banco de dados (**alteração**). *Entenderam a transação?* Prosseguindo...

Os bancos de dados tradicionais são bastante eficientes e rápidos para realizar esse tipo de transação. Inclusive, eles são muito recomendados para a galera que trata diretamente com esses dados transacionais. *Adivinha quem?* Professores e Alunos. Nós – professores – que estamos inserindo, excluindo e alterando aulas o tempo todo; e vocês – alunos – que estão acessando aulas diariamente para estudar e passar em um concurso público.

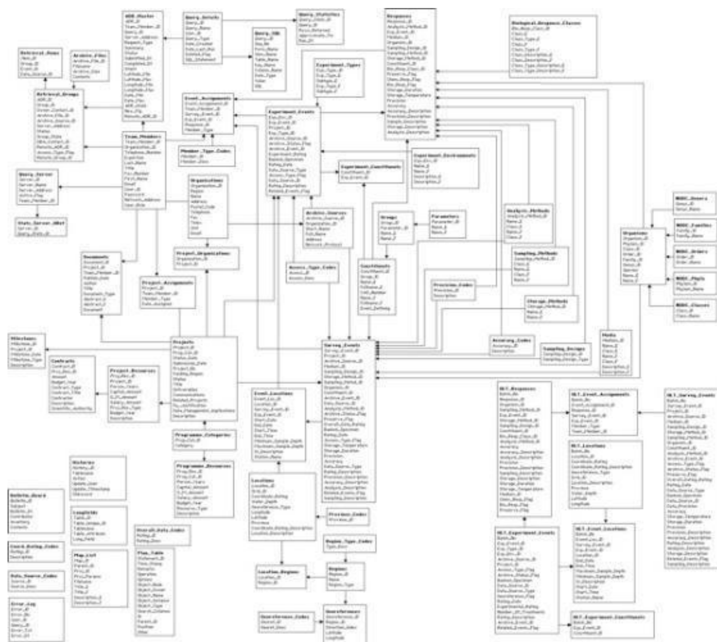


Voltando a nossa história: **vocês devem se lembrar que essa empresa de concursos começou a crescer demais e acabou contratando funcionários para diversas áreas** (Financeira, Comercial, Recursos Humanos, etc). Dito isso, eu tenho uma pergunta: *vocês acham que – para essas áreas – é mais interessante tratar de cada dado individualmente ou é mais interessante tratar mais de relatórios e consultas gerenciais?* Segunda opção!

O pessoal da **área pedagógica** precisa gerar um relatório com os professores que mais atrasam aulas ou perguntas no fórum; o pessoal da **área financeira** precisa gerar um relatório de remuneração mensal dos professores; o pessoal da **área comercial** precisa gerar um relatório de quantidade de vendas e comissões mensais; por fim, o pessoal da **área de recursos humanos** precisa gerar um relatório de cálculos trabalhistas.

Vocês conseguem entender que o intuito de gerar esses relatórios com informações consolidadas é facilitar a tomada de decisão para os gestores? Pois é, se eles possuem esse tipo de informação em mãos, eles podem tomar a decisão de – por exemplo – cobrar o professor que está atrasando as aulas; verificar quem são os professores mais bem remunerados; monitorar porque as vendas subiram ou caíram; entre outros.

No entanto, tudo tem um porém... gerar esses relatórios consolidados requer um alto poder de processamento do banco de dados! *Professor, um banco de dados transacional não é capaz de gerar esses relatórios?* Em geral, sim! Todavia, eu contei para vocês que essa empresa foi crescendo, crescendo, crescendo e foi se tornando uma gigante com milhões de alunos. O resultado disso foi que o banco de dados passou a ter mais de um bilhão de linhas...



E não é só isso! **Esse banco de dados agora possui um modelo de dados gigantesco com dezenas de relacionamentos complexos entre suas tabelas**. Como esses relatórios gerenciais necessitam de um alto poder de processamento, toda vez que eles são processados, o banco de dados fica lento para todo mundo que estiver acessando. Logo, os alunos ficam irritados porque o download das aulas está mais devagar; os professores ficam furiosos porque o upload das aulas está igual a uma tartaruga; e os funcionários ficam chateados porque todo mundo está dizendo que eles são os culpados por estarem processando esses ~~malditos~~ relatórios gerenciais. *Que pepino!*

Podemos concluir que gerar um relatório consolidado desses não é para qualquer banco de dados tradicional – nós precisamos de mais performance! *O que fazer, professor?* Bem, você pode adquirir um servidor de banco de dados transacional com o dobro do processamento, porém ele provavelmente será muito caro e, se a empresa crescer mais ainda, ele logo logo também não dará conta do recado. *Concordam comigo?*



Outra alternativa interessante seria adquirir mais um servidor de banco de dados transacional, copiar os mesmos dados do banco de dados atual, porém utilizá-lo somente para processar relatórios e, não, para processar transações. Opa, essa ideia é interessante porque o custo é

menor, no entanto **continuamos tendo que processar relatórios imensos sob demanda toda vez que for necessário – o que ainda pode ser muito lento**. Vejamos o esquema abaixo:



Legal, estamos nos aproximando da solução ideal: *que tal se nós tivéssemos mais um servidor de banco de dados, mas que já guardasse esses relatórios prontos em tabelas?* Ahhhhh... que ideia sensacional! **Galera, relatórios são complexos de gerar porque o servidor tem que processar dados de diversos relacionamentos entre dezenas ou até centenas de tabelas diferentes do banco de dados transacional e consolidá-los em um único lugar.**

Nessa solução, essas tabelas continuariam no banco de dados transacional, mas nós teríamos mais um banco de dados um pouquinho diferente chamado: **Data Warehouse (DW)**!



O Data Warehouse deixará esses relatórios já armazenados em tabelas e prontinhos para serem consultados. *Calma, professor... para que ele deixe esses relatórios armazenados em tabelas, ele terá que fazer o mesmo processamento do banco de dados tradicional!* Isso é verdade, porém o **Data Warehouse é um banco de dados especializado em gerar relatórios, logo sua performance para esse tipo de processamento é extremamente eficiente.**

Professor, de todo modo os relatórios não terão que ser alimentados? Sim, senhor! Isso é realizado por meio de um processo chamado **ETL** (**Extract**, **Transform**, **Load**). Em outras palavras, esse processo realiza três operações: **Extract** – extrai as informações do banco de dados transacional; **Transform** – transforma as informações para o formato adequado; e **Load** – carrega as informações no Data Warehouse! *Entendido?*

Esse modelo apresenta alta performance, uma vez que os usuários estão distribuídos (alunos e professores acessam o banco de dados transacional; funcionários e gestores acessam o Data Warehouse); é mais fácil realizar consultas e analisar informações, uma vez que os dados já estarão organizados em relatórios pré-definidos; o processamento é reaproveitado, isto é, uma vez que seja realizado o ETL, os relatórios estarão processados e prontos para serem consultados.

Professor, só tem vantagens? Claro que não! **Não é fácil captar e transformar os dados de fontes diversas em algo útil para o Data Warehouse – isso pode reduzir a relação**

custo/benefício; ferramentas e infraestrutura extras costumam ser necessárias, nem sempre compensando o custo de aquisição, treinamento e operação; a informação pode ficar obsoleta rapidamente; dados pouco estruturados costumam ser difíceis de se trabalhar; entre outros.

De todo modo, essa empresa era muito sinistra e continuou crescendo exponencialmente! Agora ela não é mais uma empresa – ela é uma *holding* fazendo a gestão de várias outras empresas: uma que cuida somente de concursos; outra que cuida somente de pós-graduação; outra que cuida somente de provas de certificação; outra que cuida somente de vestibulares; outra que cuida somente de um site de questões; e assim por diante.

E o número de usuários explodiu de novo porque os gestores dessas empresas também começaram a utilizar o DW! *Resultado?* Nem ele aguentou tanto acesso e começou a ficar lento de novo. *Oh, céus! De novo isso, professor?* Pois é! Nós até podemos utilizar a mesma estratégia que utilizamos na ocasião anterior, mas existe uma alternativa mais interessante: **nós podemos criar subconjuntos de um Data Warehouse**. *Como é, professor?*



Nós podemos segmentar em subconjuntos de acordo com diversos critérios, por exemplo: assuntos (Concursos, Coaching, Pós-Graduação, Certificação, Vestibulares, Questões, etc); níveis de sumarização [3] (Anual, Semestral, Trimestral, Mensal, Semanal, etc); localização (Nordeste, Norte, Centro Oeste, Sudeste, Sul). **Para finalizar, esses subconjuntos de Data Warehouse (DW) são chamados de Data Mart (DM)**.



Pronto, nós vimos uma visão geral sobre o assunto! Agora seremos mais objetivos, isso foi só para passar uma contextualização uma vez que esse é um assunto extremamente técnico.

Fechou?

Definições Básicas

INCIDÊNCIA EM PROVA: MÉDIA

Pessoal, eu dei uma visão geral sobre o tema, mas é interessante ver como são as definições acadêmicas e como isso geralmente cai em prova. Então, vejam a tabela a seguir:

Autor	Definições Acadêmicas
Ralph Kimball	**Conjunto de ferramentas e técnicas de projeto**_ [4] que – quando aplicadas às necessidades específicas dos usuários e aos bancos de dados específicos – permitirá que planejem e construam um Data Warehouse.
Bill Inmon	Coleção de dados_ [5] **orientados por assunto, integrados, variáveis com o tempo e não-voláteis**, para dar suporte ao processo de tomada de decisão.
Arun Sen	Banco de dados construídos no interesse de **suporte à decisão de negócios** e contêm dados históricos sumarizados e consolidados provenientes de registros individuais de bancos de dados operacionais.
Kenneth Laudon	Banco de dados – com ferramentas de consulta e relatório – que armazena dados atuais e históricos extraídos de vários sistemas/ambientes operacionais e **consolidados para fins de análises e relatórios administrativos**.

Vamos identificar as palavras-chaves que formam essas definições? Coleção de dados; ferramentas de consulta e relatório; orientados por assuntos; integrados; variáveis com o tempo; não-voláteis; dados históricos sumarizados e consolidados; suporte à tomada de decisão. **Notem que nós vimos basicamente todos esses conceitos em nossa contextualização**. Fiquem tranquilos porque nós vamos ver em detalhes no tópico a seguir...

Características Essenciais

INCIDÊNCIA EM PROVA: ALTA

Galera, nós vimos um contexto geral e um conjunto de definições básicas. No entanto, nós passamos direto por vários termos importantes que veremos a seguir:



- **Data Warehouse é orientado por assunto**:

Um Data Warehouse é montado por assunto, isto é, ele trata de temas específicos e importantes para o negócio da organização. Por exemplo: uma empresa do ramo de concursos pode ter como assuntos principais: concursos militares, concursos fiscais, concursos policiais, concursos de tribunais, concursos bancários, concursos legislativos, entre outros. Aqui não tem muito o que falar, essa característica é bastante intuitiva.

- **Data Warehouse é não-volátil**:

Galera, vem comigo! *Vocês sabem o que significa algo ser volátil?* **O dicionário nos diz que volátil é aquilo que muda com facilidade**. Nós vimos que os dados são carregados a partir do Banco de Dados Transacional para o Data Warehouse. *Vocês se lembram?* O lance é que o Data Warehouse é não-volátil, logo ele não muda com facilidade. Na verdade, os dados – em tese – jamais devem mudar (serem alterados ou deletados) após inseridos.

Em outras palavras, podemos dizer que os dados ficam disponíveis apenas para que os usuários realizem consultas e façam relatórios que auxiliem a tomada de decisão por parte dos gestores de uma organização. Logo, os dados disponíveis em um Data Warehouse mostram uma característica de somente leitura para os usuários finais de um banco de dados. *Essa primeira característica foi fácil, concordam?*

- **Data Warehouse é integrado**:

Professor, o que diabos é isso? **Galera, quando eu era desenvolvedor de sistemas, era comum desenvolver um software que – por exemplo – requisitava várias informações dos usuários em um formulário**. Dentre essas informações, uma muito comum perguntava em que banco o usuário possuía uma conta corrente. Como funcionava: havia um campo em branco e o usuário preenchia livremente qual era o banco em que ele tinha conta corrente.

Qual é o grande problema disso? O problema é que um usuário escrevia "*Caixa Econômica*"; outro usuário escrevia "*Caixa*"; outro usuário escrevia "*Caixa Econômica Federal*"; e – por fim – outro escrevia "*CEF*". **Galera, se os dados representam a mesma entidade, mas estão com nomes diferentes, podemos dizer que os dados não estão integrados**. Em outras palavras, ser integrado significa que os dados estão consistentes e uniformes.

Professor, uma maneira de resolver isso não seria obrigar o usuário a escolher o nome do banco em uma lista em vez de permitir que ele escreva? Booooa... é isso que nós fazemos! *No entanto, o que fazer com os dados que já foram cadastrados?* Além disso, **é comum que vários softwares diferentes, desenvolvidos por empresas diferentes, salvem dados em um mesmo banco de dados**. E se uma empresa escolheu o nome "*Caixa Econômica Federal*" e outra "*CEF*"? Já era...

Como vocês fazem para corrigir isso? Galera, nós limpamos a base de dados! *Limpamos, professor?* Sim, esse é o termo utilizado! **Isso significa que as inconsistências são removidas e os conflitos de informações são consertados antes de serem inseridos ou carregados em um Data Warehouse**! Dessa forma, podemos afirmar que os dados de um Data Warehouse são integrados de múltiplas fontes diferentes de forma consistente, padronizada e uniforme.

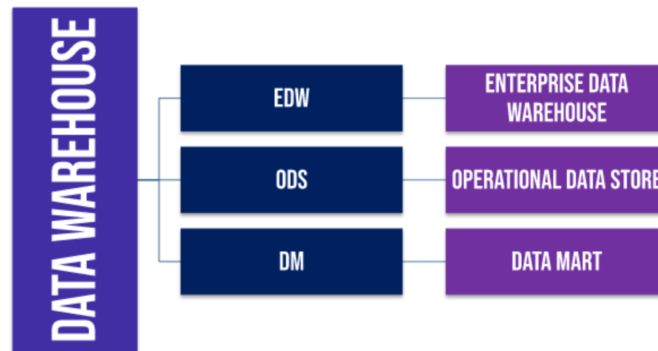
- **Data Warehouse é histórico (variável com o tempo)**:

Dados históricos são mantidos em um Data Warehouse. Por exemplo: pode-se recuperar dados de três, seis, doze meses ou até mesmo dados mais antigos de um Data Warehouse. Isso contrasta com os bancos de dados transacionais, em que geralmente apenas os dados mais recentes são mantidos. Por exemplo: um banco de dados transacional pode conter o endereço mais recente de um cliente, enquanto um DW pode conter todos os endereços associados a um cliente.

Tipos de Data Warehouse

INCIDÊNCIA EM PROVA: MÉDIA

Galera, existem basicamente três tipos de Data Warehouse: **Enterprise Data Warehouse (EDW)**, **Operational Data Store (ODS)** e **Data Mart (DM)**! *Armaria, professor... que siglas são essas?*



Vamos ver a definição dos três tipos, mas não se preocupem tanto com os dois primeiros – **nosso foco aqui é o Data Mart**. Então, vejamos abaixo:

- **Enterprise Data Warehouse (EDW):**

Traduzido como Armazém de Dados Corporativos/Abrangentes/Empresariais, é um repositório central tático/estratégico que contém todas as informações de negócio de uma organização de forma acessível para toda empresa com a finalidade de suportar as análises de decisões. Grosso modo, é possível afirmar que um EDW é um DW de larga escala – grande e abrangente – que é utilizado pela organização como um todo e que integra informações de diversas fontes de dados.

- **Operational Data Store (ODS):**

Traduzido como Armazenamento de Dados Operacionais, trata-se de um repositório intermediário de dados utilizado para relatórios operacionais e como uma fonte de dados para o Enterprise Data Warehouse (EDW). É um elemento complementar a um EDW em um cenário de suporte a decisões, e é usado para relatórios, controles e tomada de decisões operacionais – em oposição ao EDW, que é utilizado para suporte de decisões táticas e estratégicas.

- **Data Mart (DM):**

Traduzido como Mercado de Dados, trata-se de um subconjunto de dados de um Data Warehouse. Geralmente são dados referentes a um assunto em especial (Ex: Vendas, Estoque, Controladoria, entre outros) ou diferentes níveis de sumarização (Ex: Vendas Anual, Vendas Semestral, Vendas Mensal, entre outros), que focalizam uma ou mais áreas específicas. Seus dados são obtidos do DW de forma indexada para suportar intensas pesquisas.

O Data Mart é considerado um pequeno Data Warehouse – ou um subconjunto/segmentação de um Data Warehouse – que contém dados de uma organização em geral restritos a uma unidade de negócio ou a um processo de negócio específico com a finalidade de dar suporte ao processo decisório da alta direção por meio da criação de uma visão personalizada dos

dados. **De acordo com Date, trata-se de um repositório especializado e volátil (mas é o único autor que diz isso)!**

Data Marts geralmente lidam com um único assunto, sendo frequentemente construídos e gerenciados por um único departamento comercial. Eles normalmente obtêm dados de apenas um pequeno número de fontes e – como consequência – são geralmente menos complexos, mais econômicos, mais fáceis de construir e mais fáceis de manter do que Data Warehouses. Vejamos algumas de suas principais características:

- Um Data Mart permite um acesso mais rápido aos dados, recuperando um conjunto específico de dados para BI e Relatórios. Como resultado, ajuda a acelerar os processos de negócios;

- Sendo focado em um assunto, é mais fácil implementar um Data Mart, sendo mais econômico em comparação com a construção de uma empresa Data Warehouse;

- O uso de um Data Mart é fácil, visto que ele é projetado de acordo com os requisitos de um grupo específico de usuários que trabalham em um departamento específico;

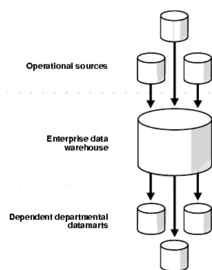
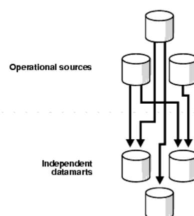
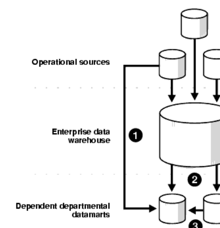
- Um Data Mart é comparativamente mais adaptável que um Data Warehouse. Qualquer alteração no modelo de dados pode ser fácil e rapidamente incorporada no Data Mart;

- Em um Data Mart, os dados são particionados e segmentados, o que permite um controle de acesso em um nível mais granular.

Os Data Marts são utilizados por uma única unidade de negócios para armazenar suas informações. **Digamos que o departamento de vendas de uma empresa tenha que armazenar seus dados de negócios – eles poderão utilizar um Data Mart para armazenar seus dados.** Quando as informações precisam ser visualizadas pelos superiores, elas são carregadas em um Data Warehouse e são, então, transformadas em insights usando software de Business Intelligence.

Tipos de Data Mart	Descrição
Dependente	Nesse caso, um conjunto de fontes de dados operacionais são integradas em um EDW, que – por sua vez – são divididos em diversos Data Marts

	dependentes.
Independente	Nesse caso, um conjunto de fontes de dados operacionais permitem gerar um ou mais Data Marts independentes, sem a necessidade de criar um EDW.
Hhíbrido	Nesse caso, um conjunto de fontes de dados operacionais podem ser tanto integradas em um EDW para posteriormente permitir a criação de Data Marts quanto podem gerá-los de maneira independente.

Dependente	Independente	Híbrido
		

#VamosPraticar#

(TCE/PA – 2016)

Um Data Mart é:

a) uma técnica de mineração de dados para a extração de conhecimento das bases de dados de uma organização.

b) um grande data warehouse que consolida todas as informações de uma organização a serem tratadas e organizadas.

c) um sistema para a monitoração e processamento das funções básicas e rotineiras de uma organização.

d) um conjunto de metadados para a análise e monitoração de um data warehouse organizacional.

e) um pequeno data warehouse que dá suporte à decisão para uma específica unidade de negócio.

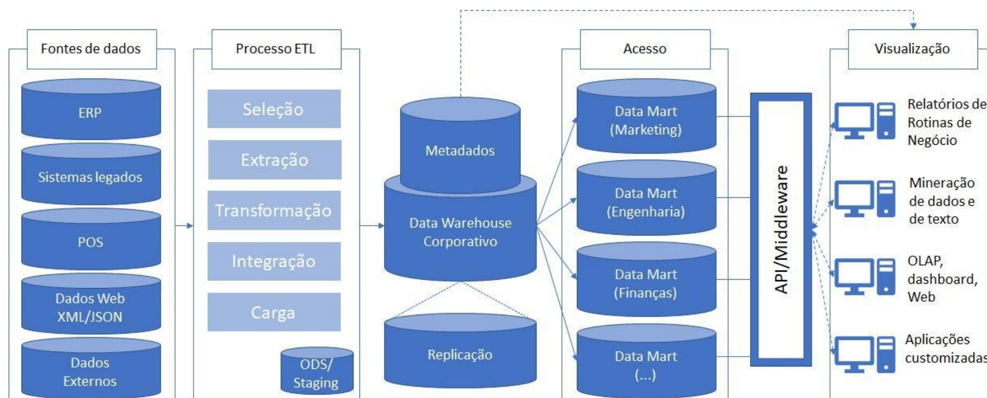
Comentários: é um pequeno Data Warehouse que dá suporte à decisão para uma específica unidade de negócio (Letra E).

#VamosPraticar#

Processo de Data Warehouse

INCIDÊNCIA EM PROVA: BAIXA

Galera, nós já vimos componentes suficientes para finalmente entender como funciona o processo de um Data Warehouse. Vejam a figura abaixo:



Vamos por partes: à esquerda, temos os Data Sources! *O que é isso, Diego?* **Cara, essas são as fontes de dados**. *Vocês se lembram que nós vimos que um DW possui dados de diversas fontes distintas?* Pois é, essa imagem mostra alguns exemplos! Notem que uma fonte de dados pode ser um ERP (Enterprise Resource Planning), que nada mais é que um sistema de informação que integra dados e processos de uma organização. *O que mais?*

Sistemas Legados, que são aqueles sistemas beeeeem antigos de uma organização (em geral, com tecnologias ultrapassadas) que ainda fornecem serviços de negócio essenciais. *Eles também servem de fontes de dados para um Data Warehouse?* Claro, uma fonte importantíssima! *Point of Sale* (POS), que **são máquinas registradoras ou de cartão de crédito que realizam transações de venda – em determinados contextos, também podem ser importantes fontes de dados**.

Outra fonte de dados são aquelas advindas de aplicações web e bancos de dados transacionais – além de quaisquer outras fontes externas de dados. Pessoal, esses são só exemplos típicos de fontes de dados, mas podem existir dezenas de outras. **De todo modo, a imagem quer mostrar que diversas fontes de dados servem de entrada para um Enterprise Data Warehouse (EDW)**. No entanto, notem que – antes de serem carregados – eles devem passar pelo processo de ETL.

Vocês se lembram que se trata do processo de ETL – Extract, Transform e Load (Extração, Transformação e Carga)? Essas são as atividades principais, porém existem outras subdivisões: primeiro, selecionam-se os dados que se deseja carregar; segundo, extrai esses dados de suas fontes para uma *Stage Area*, **que nada mais é que a localização temporária de onde os dados dos sistemas de origem são copiados**.

Em seguida, os dados são transformados, isto é, uma série de regras ou funções são aplicadas aos dados extraídos para prepará-lo para carregar em seu destino final; depois os dados são integrados, ou seja, são uniformizados em um padrão específico; por fim, os dados são finalmente carregados no nosso armazém de dados empresarial. **Agora notem que existem ainda dois conceitos: metadados e replicação**.

Metadados são dados sobre o próprio Enterprise Data Warehouse, por exemplo: estruturas e atributos de tabelas, especificações do modelo de dados, rotinas de acesso, log de extrações, entre outros. **Notem que os metadados auxiliam na criação da visualização dos dados por meio de diversas ferramentas.** Para finalizar, Replicação é uma cópia que pode ser realizada do próprio Enterprise Data Warehouse para diversas funcionalidades.

Agora notem que ele pode ser dividido em vários Data Marts: Marketing, Engenharia, Finanças, entre outros. Para finalizar, notem que usuários e aplicações utilizam middlewares para acessar Data Marts ou Data Warehouses – eles não acessam os dados diretamente. Fim! **A partir daí, os dados estarão disponíveis para serem analisados e você poderá descobrir informações relevantes para o negócio de uma organização**.

Kimball x Inmon

INCIDÊNCIA EM PROVA: MÉDIA

Pessoal, deixa eu contar uma coisa para vocês: quando o assunto é Data Warehouse, não tem para ninguém – existem dois nomes que comandam a bibliografia! Ei-los...



Galera... esses caras são tão sinistros que algumas provas cobram o nome deles explicitamente! *Professor, por que tudo isso?* Porque ambos foram pioneiros na teoria sobre Data Warehouse no início da década de noventa, mas não só por isso! **Esses dois caras concordam em vááárias coisas sobre como implementar um Data Warehouse, no entanto há alguns pontos em que eles discordam diametralmente**. Vamos começar pela parte que eles concordam...

A Arquitetura de Inmon é chamada de Enterprise Data Warehouse (sim, aquele que nós vimos lá atrás) e a Arquitetura de Kimball é chamada de Dimensional Data Warehouse. Via de regra, elas são muito parecidas contendo um único repositório integrado de dados corporativos que suporta a análise de informações em toda a organização. Por outro lado, esses discordam veementemente na forma como um Data Warehouse deve ser construído. Vejamos:

De maneira simplista, pode-se afirmar que Inmon defende uma abordagem de cima para baixo (*top-down*) e Kimball defende uma abordagem de baixo para cima (*bottom-up*). Em outras palavras, Inmon acha que a maneira correta é construir o Data Warehouse e depois dividi-lo em Data Marts; já Kimball acha que a maneira correta é construir vários Data Marts e, só então, integrá-los em um Data Warehouse. Em suma:

- **Abordagem de Inmon:**

Recomenda a construção do DW que siga a abordagem de cima para baixo. **Na filosofia de Inmon, é começando com a construção de um grande Enterprise Data Warehouse centralizado**, onde todos os dados disponíveis dos sistemas transacionais são consolidados em uma coleção de dados orientada ao assunto, integrada, variante no tempo e não volátil que suporta a tomada de decisões. Em seguida, os Data Marts são construídos para necessidades analíticas de departamentos.

- **Abordagem de Kimball:**

Recomenda a construção do DW que segue a abordagem de baixo para cima. **Na filosofia de Kimball, inicia-se primeiro com Data Marts de missão crítica que atendem às necessidades analíticas dos departamentos**. Em seguida, integram-se esses Data Marts para consistência de dados por meio do chamado barramento de informações. Kimball faz uso do modelo dimensional para atender às necessidades dos departamentos em várias áreas da empresa.

Vamos detalhar um pouco mais a abordagem de Kimball! Esse autor dividia sua abordagem basicamente em quatro etapas apresentadas a seguir:

Etapa	Nome	Descrição
1	**Identificar os Processos de Negócio**	Processos de negócio são atividades de negócio realizadas por uma organização, como receber pedidos, realizar faturamento, receber pagamentos, atender chamadas de serviço, cadastrar alunos, realizar um procedimento médico ou processar sinistros. Ouvir os usuários é o meio mais eficiente para selecionar um processo de negócio. As medições de desempenho que eles analisam no Data Warehouse resultam de processos de medições do negócio.

2	**Identificar a Granularidade**	Identificar a granularidade significa especificar exatamente o que uma linha específica da tabela de fatos representa. A granularidade exprime o nível de detalhe associado às medições da tabela de fatos. Ela fornece a resposta para a pergunta: "*Como você descreve uma única linha na tabela de fatos?*". Em suma, identificar a granularidade é identificar o nível de detalhe de um processo de negócio.
3	**Identificar as Dimensões**	Você precisa rodear a tabelas de fatos com um conjunto robusto de dimensões que representem todas as descrições possíveis que assumem valores únicos no contexto de cada medição. As dimensões representam o "quem, o que, onde, quando, por que e como" associados ao evento. Exemplo: data, produto, cliente, funcionário e local.
4	**Identificar os Fatos**	Fatos são métricas resultantes do processo de negócio, isto é, quantidades, valores, medidas, entre outros. Os fatos são determinados pela resposta à pergunta: *O que estamos medindo?* Os fatos também podem ser derivados ou calculados a partir de outras medidas. Os usuários de negócios estão profundamente interessados em analisar essas métricas de desempenho.

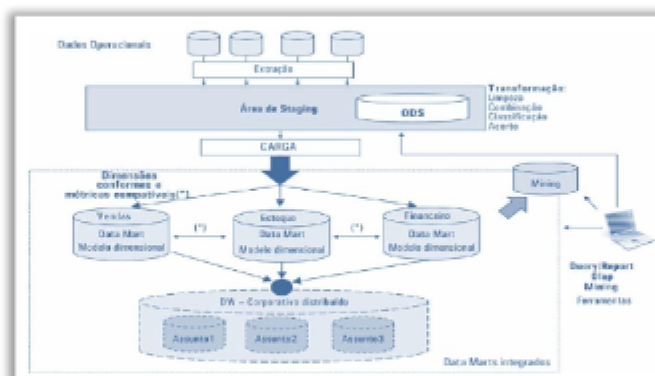
Característica	Kimball (DM)	Inmon (EDW)
Abordagem	Bottom-Up (Data Marts Data Warehouse)	Top-Down (Data Warehouse Data Marts)
Orientação	Orientado a processos de negócio	Orientado a dados
investimento Inicial	Baixo	Alto
Normalização	Desnormalizado	Normalizado (3ª Forma Normal)
Direcionamento	Usuários finais	Profissionais de TI
Consultas	Realizadas no Data Warehouse	Realizadas no Data Mart

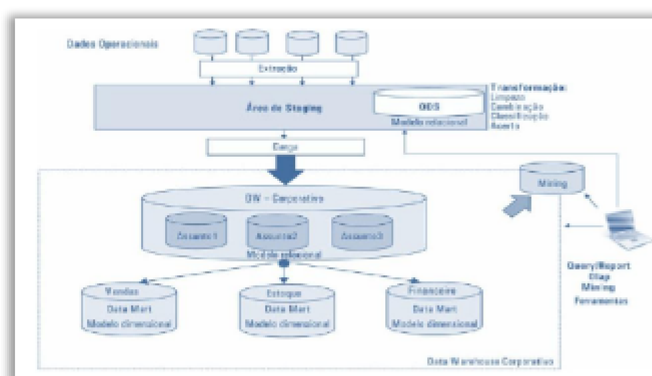
Antes de prosseguir, é importante explicar duas palavrinhas: normalizado e multidimensional! Explicando beeem rapidamente, a normalização é um conjunto de regras aplicadas a um banco de dados para reduzir a redundância, aumentar a integridade e melhorar o desempenho (exceto de consultas). Como mostra a tabela, o banco de dados do Kimball não se importa com dados redundantes, já o banco de dados de Inmon se importa (e é normalizado).

Agora vocês se lembram que temos a modelagem relacional? **Temos também a modelagem multidimensional – em geral, utilizada em Data Warehouses! A Modelagem dimensional é uma técnica normalmente usada com D [6]ata Warehouses**. Segundo Kimball [7], a modelagem dimensional é a única técnica viável para bancos de dados que devem responder consultas em um DW, uma vez que a modelagem relacional é mais útil para representação de transações.

Não há certo ou errado entre essas duas ideias, uma vez que elas representam apenas diferentes filosofias de armazenamento de dados. Na realidade, a maioria das empresas adota a abordagem de Kimball, porque a maioria dos Data Warehouses começam como um esforço departamental e, portanto, originaram-se como um Data Mart. Somente quando mais Data Marts são construídos posteriormente é que eles evoluem para um Data Warehouse.

Por fim, eu deixo um desafio para vocês: nos diagramas abaixo, um representa a abordagem de Kimball e outro representa a abordagem de Inmon. *E aí, qual que é qual?*





ETL

INCIDÊNCIA EM PROVA: MÉDIA

Galera, você precisa carregar seu Data Warehouse regularmente para que possa servir ao propósito de facilitar a análise de negócios. **Para fazer isso, os dados de um ou mais sistemas/ambientes operacionais precisam ser extraídos e copiados no Data Warehouse**. O desafio nos ambientes de Data Warehouse é integrar, reorganizar e consolidar grandes volumes de dados de muitos sistemas, fornecendo assim uma nova base de informações unificada para análise de dados.

O processo de extrair dados de sistemas de origem e trazê-los para o Data Warehouse é comumente chamado de ETL (*Extract*, *Transform* e *Load*). **Galera, é importante entender que o ETL se refere a um processo mais amplo e, não, a três etapas bem definidas**. A verdade é que a sigla ETL é simplista demais, uma vez que omite algumas fases e implica que cada uma das outras fases do processo é distinta. No entanto, o processo como um todo é conhecido como ETL!

Ele é o processo mais crítico e demorado na construção de um Data Warehouse, uma vez que consiste na extração dos dados de fontes homogêneas ou heterogêneas; na transformação e limpeza destes dados; e na carga dos dados no DW. Pessoal, as decisões estratégicas – aquelas mais importantes de uma organização – são tomadas com base nas informações geradas através dos dados armazenados no Data Warehouse.

Se estes dados não forem corretamente trabalhados no processo de extração, as informações geradas através deles farão com que decisões sejam tomadas erroneamente, podendo afetar diretamente os negócios da organização. *Já viram a responsabilidade, não é?* Galera, hoje em dia todo órgão público de médio/grande porte possui – em sua área de tecnologia da informação – uma gerência responsável por toda essa parte de *Business Intelligence*.

Essa área geralmente é responsável pela criação de Data Warehouses e pelo Processo de ETL! Aliás, não sei se vocês sabem, mas a maior parte do esforço exigido no desenvolvimento de um DW é consumido neste momento – eu diria que cerca de uns 80%! **Logo, ele tem que ser bem feito e refletir a realidade de modo que analistas consigam tirar informações relevantes para o processo de tomada de decisões gerenciais.**

Sabe outra curiosidade? Eu disse que 80% do esforço de desenvolvimento de um DW é ETL! *Sabia que 60% do Processo de ETL é só com a Extração?* Pois é, essa etapa deve se basear na busca das informações mais importantes em aplicações ou fontes externas. Pessoal, essa busca geralmente não é tão simples! **Dados podem estar distribuídos em dezenas de lugares, tecnologias ou plataformas diferentes necessitando de formas de extração distintas para cada local.**

É, galera – não é tudo mil maravilhas! Você perde muito tempo durante a extração porque frequentemente é necessário implementar diversas formas diferentes de extrair dados de seus locais de origem. E sem carga de dados, não temos Data Warehouse! **No momento da criação de um DW, é comum realizar uma carga de dados inicial, mas com o decorrer do tempo a extração deve estar preparada apenas para fazer cargas incrementais.**

A carga incremental é muito mais eficiente visto que carrega apenas os dados que foram inseridos desde a carga inicial. *Professor, e para onde vão esses dados que foram extraídos, mas ainda não foram transformados e carregados no Data Warehouse?* Excelente pergunta! Resposta: **eles vão para um local chamado *Stage Area***! *O que é isso, professor?* Vocês devem se lembrar que uma das características essenciais de um Data Warehouse é sua integridade.

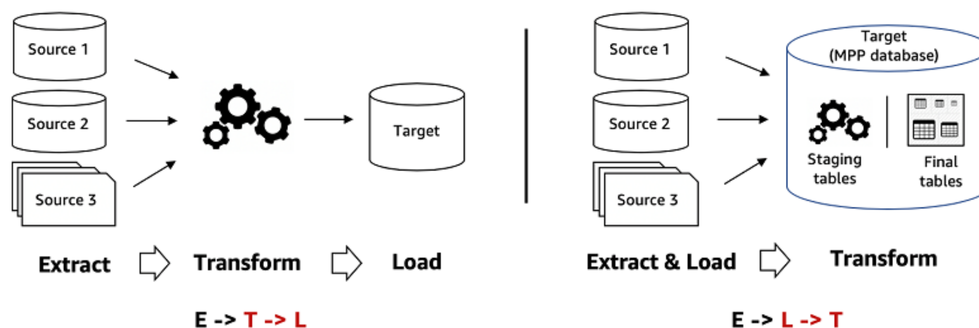
Relembrando: os dados devem estar padronizados, consistentes e uniformizados antes de serem carregados no Data Warehouse! Para tal, eles precisam ser limpos, entre outras

atividades. **Isso tudo ocorre na Stage Area, portanto ela é basicamente uma localização temporária de onde os dados de diversas fontes de origem são copiados**. *Sabe quando você copia e cola algo em um sistema operacional?* O que você copiou vai para um local chamada Área de Transferência!

Adivinhem só: *Stage Area* é similar à Área de Transferência, isto é, uma localização de armazenamento temporário. **Seguindo: a transformação dos dados é a fase subsequente à extração. Esta fase não só transforma os dados, mas também realiza a limpeza dos mesmos**. A correção de erros de digitação, a descoberta de violações de integridade, a substituição de caracteres desconhecidos, a padronização de abreviações podem ser exemplos dessa limpeza.

Já a fase de carga consiste no carregamento dos dados para seu destino final. Esse processo pode variar muito, uma vez que alguns Data Warehouses podem adicionar as informações existentes mensalmente (com dados cumulativos e atualizados) ao passo que outros podem adicionar dados a cada hora, dia ou semana. Galera, era isso que era cobrado até pouco tempo atrás, mas as coisas mudaram... agora estão cobrando até o ELT!

Professor, você errou a sigla! Errei, não... é isso mesmo! ELT é uma inversão da ordem das etapas que vimos anteriormente: em vez de fazermos (1) Extração, depois (2) Transformação e só depois a (3) Carga, nós fazemos a (1) Extração, depois a (2) Carga e só depois a (3) Transformação dos dados. **Primeiramente, é importante destacar que os processos de ETL e ELT – apesar de diferentes – solucionam o mesmo problema.**



O ELT é um processo mais ágil para o carregamento e o processamento de dados, visto inverte a ordem das etapas de transformação de dados da abordagem tradicional de ETL. Nesse cenário, a transformação de dados ocorre logo após a coleta e o carregamento das informações em um repositório de dados centralizado e, não, antes. Com isso, é possível transformar dados brutos em dados modelados dentro de um Data Warehouse.

Essa simples inversão de etapas **reduz consideravelmente o tempo de carregamento de dados**, permitindo que a transformação de dados seja feita por analistas de dados comuns, sem a dependência de profissionais altamente técnicos como desenvolvedores e engenheiros de dados. Na prática, os engenheiros de dados se focam apenas nas etapas de extração e carregamento, atribuições características de sua área de atuação.

Dessa forma, **a responsabilidade da transformação de dados fica nas mãos de profissionais próximos à empresa que conhecem as regras de negócio**, como analistas e cientistas de dados, trazendo benefícios como: menor tempo gasto entre extração e transformação de dados; maior eficiência na implementação de projetos; menor dependência de profissionais altamente técnicos; e independência dos analistas para o manuseio de dados brutos.

Quando a ETL foi criada, o armazenamento de dados em repositórios era extremamente caro. Além disso, o processamento dessas informações era extremamente lento e, por isso, pouquíssimas organizações podiam investir nessas tecnologias. **Ocorre que essa realidade mudou com a computação na nuvem (cloud computing).** Hoje, o que era lento tornou-se rápido; e o que era caro tornou-se acessível.

Assim, considerando o volume massivo de dados para processamento nas operações modernas, muitas empresas têm optado pelo ELT em vez do tradicional ETL, buscando mais flexibilidade e agilidade nos seus pipelines de dados. **O ELT permite analisar grandes volumes de dados de forma rápida, com menos manutenção e gastos de recursos – além de possibilitar uma maior colaboração entre equipes de negócio e profissionais técnicos.**

Por essa razão, ainda que o ELT seja um processo mais complexo, exigindo mais esforços nos quesitos de design e infraestrutura do que o ETL, a primeira oferece mais vantagens no longo prazo. **Além disso, o ELT leva menos tempo no carregamento e na transformação de dados e, por isso, é mais econômico.** No entanto, caso o repositório de dados de destino não seja robusto para comportar um volume massivo de dados, o processo de ETL pode ser uma opção mais viável.

Bem... agora que as provas começaram a cobrar o ELT, logo 99% das questões ainda serão sobre ETL! De toda forma, vamos resumir as diferenças na tabelinha a seguir:

Critérios	ETL	ELT
Tempo de Carga	Uso de sistemas distintos implica uma maior demora/latência para o carregamento de dados.	Uso de um sistema de carregamento integrado implica uma menor demora/latência para o carregamento de dados
Tempo de Transformação	Demora considerável – particularmente na transformação de grandes volumes de dados. Além disso, à medida que os volumes crescem, o tempo de transformação aumenta.	Transformação de dados feita de maneira integrada, visto que sua velocidade independe do tamanho ou da complexidade dos dados – graças às tecnologias de infraestrutura na nuvem.

Custo de Manutenção	Altos custos de manutenção, dado que a atualização do repositório de dados requer o trabalho recorrente de profissionais caros e escassos no mercado, como engenheiros de dados e desenvolvedores.	Baixos custos de manutenção, pois os dados estão sempre disponíveis e acessíveis para todos sem a necessidade de profissionais extremamente especializados.
Complexidade de Implementação	Requer menos conhecimento aprofundado de projeto e arquitetura do repositório de dados.	Requer conhecimento profundo de ferramentas avançadas de análise e uma arquitetura bem estruturada do repositório de dados.
Limitação de dados	Dados devem ser selecionados a priori, ou seja, o que não for selecionado é perdido.	Dados brutos ficam disponíveis, por isso a retenção de dados depende apenas da política de retenção eleita pela operação de dados.
Suporte para Banco de Dados	ETL foi desenvolvido para dar suporte a banco de dados relacionais, locais e sistemas legados.	ELT foi construído para suportar grandes volumes e diversas fontes de dados, estruturadas e não estruturadas, de forma escalável em infraestruturas de nuvem.
Usabilidade	Tabelas fixas, cronogramas fixos, predominantemente utilizados por profissionais técnicos de TI, desenvolvedores e engenheiros de dados.	Ad hoc, abordagem escalável, flexível e colaborativa, utilizável tanto por profissionais técnicos quanto pelo usuário final de negócio.
Custo / Benefício	Abordagem não econômica, especialmente para empresas de pequeno e médio portes.	Por ser escalável, adaptável e acessível para negócios de todos os tamanhos, é uma solução

		muito mais viável, econômica e moderna.
Dados Sensíveis	É mais adequado para o armazenamento de dados sensíveis e confidenciais, dado que permite a transformação antes da carga (Ex: anonimização dos dados).	É menos adequado para o armazenamento de dados sensíveis e confidenciais, dado que a transformação ocorre somente após a carga.
Repositório	Comumente utilizado em Armazém de Dados (Data Warehouses).	Comumente utilizado em Lagos de Dados (Data Lakes).

É importante mencionar que esses sistemas são complementares: não necessariamente é preciso fazer a substituição de um por outro. Pode-se afirmar que o Data Lake é ótimo para realizar análises de Machine Learning/Deep Learning, Data Discovery e Análises Preditivas; já o Data Warehouse é bastante útil para análises de Business Intelligence, Visualização de Dados, Criação de Relatórios e Análises Transacionais.

12.3 Análise de Informações - BI e Data Warehouse - Resumo

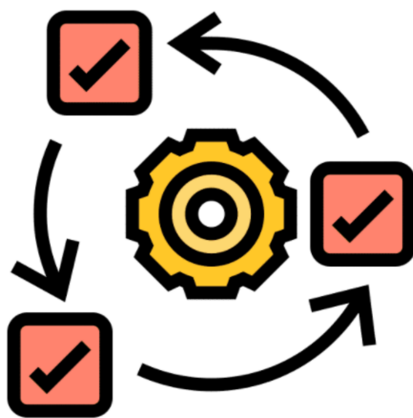
Resumo

BUSINESS INTELLIGENCE
Business Intelligence (Inteligência de Negócio) é um termo criado pelo Gartner Group utilizado para representar um conjunto de processos, técnicas, metodologias, habilidades, ferramentas e capacidades utilizadas para acessar, coletar, organizar, tratar, analisar, cruzar, processar, compartilhar e monitorar dados de diversas fontes com o intuito de gerar informações e relatórios analíticos que suportem a gestão corporativa, a definição de estratégias e a tomada de decisão em ambientes de negócio.

Tipos de Análises	Descrição

Análise Descritiva	É o exame de dados ou conteúdos para responder à pergunta: "O que aconteceu?" ou "o que está acontecendo?". É caracterizado pelo Business Intelligence tradicional e visualizações como gráficos de pizza, gráficos de barras, etc que servem para que gestores tomem decisões.
Análise Diagnóstica	É uma forma de análise avançada que examina dados ou conteúdos para responder à pergunta: "Por que isso aconteceu?", e é caracterizada por técnicas como detalhamento, descoberta de dados, mineração de dados e correlações. Esse é o tipo de análise empregada em investigações de causa-raiz.
Análise Preditiva	Combina técnicas de estatística, mineração de dados e aprendizagem de máquina para encontrar significado em grandes quantidades de dados, trabalhando com probabilidades, predições, entre outros para antecipar comportamentos futuros com base em eventos passados. Responde à pergunta: "O que vai acontecer?";
Análise Prescritiva	É uma forma de análise avançada que examina os dados ou os conteúdos para responder à pergunta: "O que deve ser feito?" ou "O que podemos fazer caso algo aconteça?", e é caracterizada por técnicas como análise de gráficos, simulação, redes neurais, mecanismos de recomendação, heurísticas, aprendizado de máquina, etc.

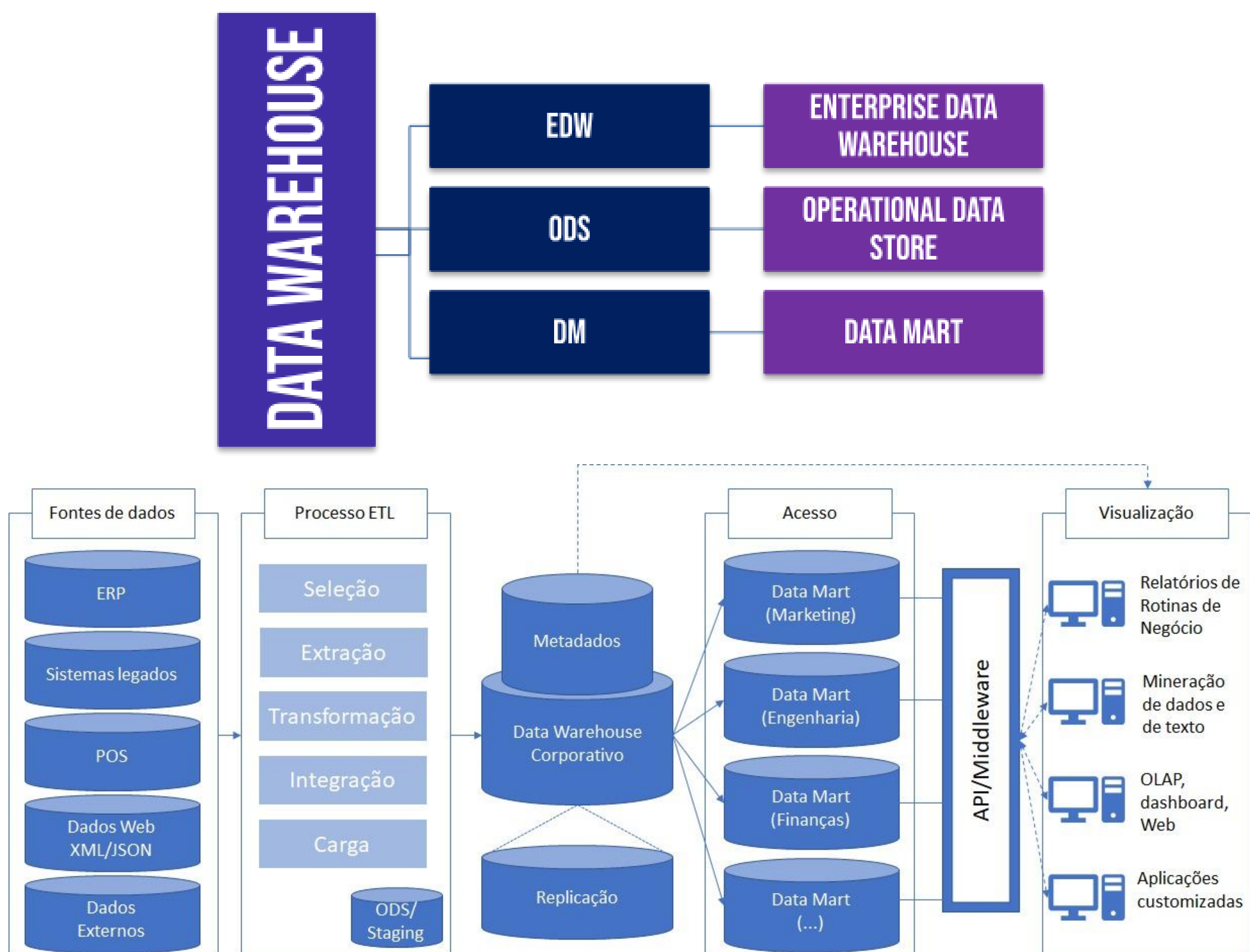
ARQUITETURA BI	Data Warehouse	**Trata-se de uma coleção de dados orientados por assunto, integrados, variáveis com o tempo e não-voláteis, para dar suporte ao processo de tomada de decisão.**
		TRATA-SE DE um conjunto de ferramentas para manipular, minerar e analisar dados, criar relatórios, realizar consultas sob demanda, entre outros.

		Trata-se de um **conjunto de ferramentas que permitem comparar e compartilhar o desempenho de negócios de uma organização** em relação às metas estabelecidas.
		Trata-se de um **conjunto de ferramentas que permitem a rápida e abrangente visualização de dados** por meio de dashboards e softwares similares.

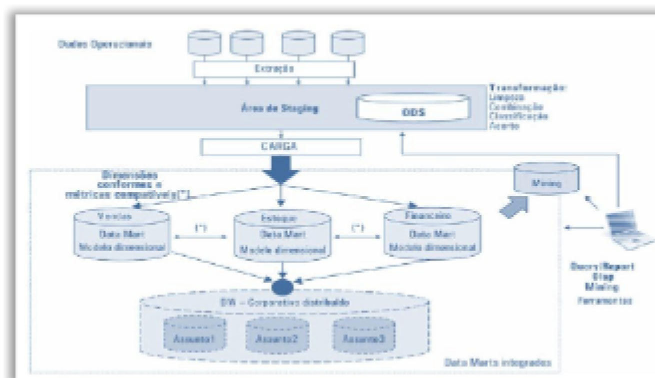
AUTOR	DEFINIÇÕES ACADÊMICAS de data warehouse
Ralph Kimball	**Conjunto de ferramentas e técnicas de projeto** que – quando aplicadas às necessidades específicas dos usuários e aos bancos de dados específicos – permitirá que planejem e construam um Data Warehouse.
Bill Inmon	Coleção de dados **orientados por assunto, integrados, variáveis com o tempo e não-voláteis**, para dar suporte ao processo de tomada de decisão.
Arun Sen	Banco de dados construídos no interesse de **suporte à decisão de negócios** e contêm dados históricos sumarizados e consolidados provenientes de registros individuais de bancos de dados operacionais.
Kenneth Laudon	Banco de dados – com ferramentas de consulta e relatório – que armazena dados atuais e históricos extraídos de vários sistemas/ambientes operacionais e **consolidados para fins de análises e relatórios administrativos**.

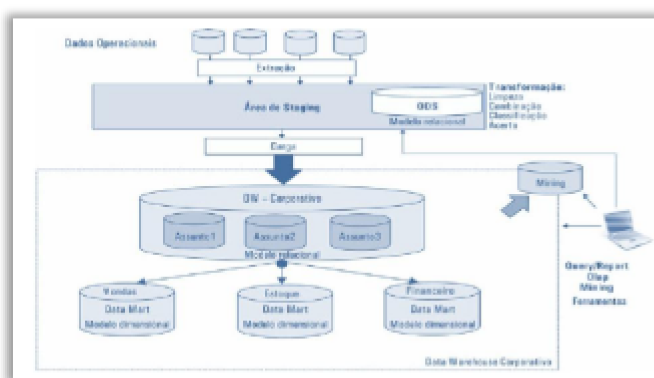


ENTERPRISE DATA WAREHOUSE (EDW)	Traduzido como Armazém de Dados Corporativos ou Abrangentes, é um repositório central tático/estratégico que contém todas as informações de negócio de uma organização de forma acessível para toda empresa com a finalidade de suportar as análises de decisões. Grosso modo é possível afirmar que um EDW é um DW de larga escala – grande e abrangente – que é utilizado pela organização como um todo e que congrega informações de diversas fontes de dados.
OPERATIONAL DATA STORE (ODS)	Traduzido como Armazenamento de Dados Operacionais, trata-se de um repositório intermediário de dados utilizado para relatórios operacionais e como uma fonte de dados para o Enterprise Data Warehouse (EDW). É um elemento complementar a um EDW em um cenário de suporte a decisões, e é usado para relatórios, controles e tomada de decisões operacionais – em oposição ao EDW, que é utilizado para suporte de decisões táticas e estratégicas.
DATA MART (DM)	Traduzido como Mercado de Dados, trata-se de um subconjunto de dados de um Data Warehouse. Geralmente são dados referentes a um assunto em especial (Ex: Vendas, Estoque, Controladoria, entre outros) ou diferentes níveis de sumarização (Ex: Vendas Anual, Vendas Semestral, Vendas Mensal, entre outros), que focalizam uma ou mais áreas específicas. Seus dados são obtidos do DW – indexados para suportar intensas pesquisas.



CARACTERÍSTICAS	KIMBALL (DM)	INMON (EDW)
Abordagem	Bottom-Up (Data Marts Data Warehouse)	Top-Down (Data Warehouse Data Marts)
Orientação	Orientado a processos de negócio	Orientado a dados
investimento Inicial	Baixo	Alto
Normalização	Desnormalizado	Normalizado (3ª Forma Normal)
Direcionamento	Usuários finais	Profissionais de TI
Consultas	Realizadas no Data Warehouse	Realizadas no Data Mart





12.4 Mapa Mental - Análise de Informações - BI e Data Warehouse

Mapa Mental







12.5 Questões Comentadas - Análise de Informações - BI e Data Warehouse - Multibancas

#VamosPraticar#

Questões Comentadas – Cespe

26. (CESPE / SECONT/ES - 2018)

Data Warehouse é uma coleção de dados orientados ao assunto, que tem como características visão conceitual multidimensional, operações interdimensionais irrestritas, dimensões e níveis de agregação ilimitados, sendo que os dados são não voláteis e variantes no tempo.

Comentários:

Data Warehouse é uma coleção de dados orientados por assunto? Sim, de acordo com Bill Inmon. *Ele tem como característica visão conceitual multidimensional?* Sim, Data Warehouse utiliza uma modelagem multidimensional e, não, modelo entidade-relacionamento. *Realiza operações interdimensionais irrestritas?* Sim, há infinitas possibilidades. *Dimensões e níveis de agregação são ilimitados?* Sim, você possui diversos níveis de granularidade. Por fim, os dados são não-voláteis e variantes no tempo, de acordo com Bill Inmon.

Gabarito: Correto

27. (CESPE / TJ/SE - 2018)

Os dados armazenados em um *DataWarehouse* devem estar integrados, temporalmente identificados, orientados ao assunto e devem ser protegidos, de modo a se evitar o acesso do usuário.

Comentários:

De acordo com Bill Inmon, um Data Warehouse é uma coleção de dados orientados por assunto, integrados, variáveis com o tempo e não-voláteis, para dar suporte ao processo de tomada de decisão. No entanto, **os dados devem estar protegidos para evitar o acesso do usuário não autorizado**, mas devem estar acessíveis aos usuários autorizados.

Gabarito: Errado

58. (CESPE / INMETRO – 2010)

Assinale a opção correta com relação aos conceitos de *data mart*.

a) O data warehouse é constituído de apenas um data mart.

b) O data mart pode ser visto como a restrição do data warehouse a um único processo de negócios.

c) Todo data mart deve ser representado por um modelo orientado a objetos.

d) O data mart possui uma estrutura diferente do data warehouse.

e) O que diferencia um data mart de um data warehouse é a forma de acesso aos dados.

Comentários:

(a) Errado. Um Data Warehouse pode ser constituído de diversos Data Marts; (b) Correto. Pode ser uma restrição a uma unidade, assunto ou processo de negócio específico; (c) Errado. Em geral, eles utilizam modelos multidimensionais; (d) Errado. Data Mart e Data Warehouse possuem estruturas semelhantes; (e) Errado. A forma de acesso aos dados também é semelhante.

Gabarito: Letra B

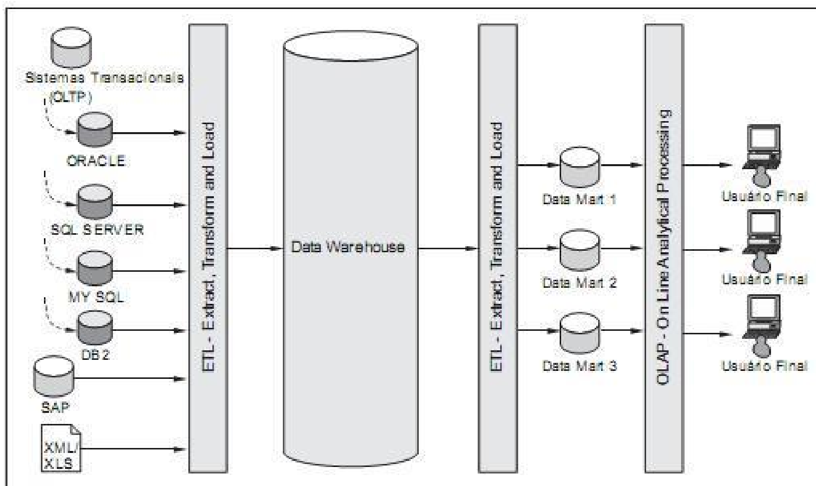
#VamosPraticar#

#VamosPraticar#

Questões Comentadas – FCC

71. (FCC / AL - MS – 2016)

Considere que a equipe de Analistas de Desenvolvimento de Sistemas do CNMP está projetando a arquitetura para o Data Warehouse (DW) da instituição, conforme mostra a figura abaixo:



QuestoesdeCONCURSOS.com.br

É correto afirmar que esta arquitetura:

a) é *bottom-up*. Garante a existência de um único conjunto de aplicações para ETL, ou seja, extração, limpeza e integração dos dados, embora os processos de manutenção e monitoração fiquem descentralizados.

b) é *bottom-up*, pois primeiro a equipe cria um DW e depois parte para a segmentação, ou seja, divide o DW em áreas menores gerando pequenos bancos orientados por assuntos aos departamentos.

c) é *bottom-up*. Permite um rápido desenvolvimento, pois a construção dos *Data Marts* é altamente direcionada. Normalmente um *Data Mart* pode ser colocado em produção em um período de 2 a 3 meses.

d) é *top-down*. A partir do DW são extraídos os dados e metadados para os *Data Marts*. Nos *Data Marts* as informações estão em maior nível de sumarização e, normalmente, não apresentam o nível histórico encontrado no DW.

e) é *top-down*, pois possui um retorno de investimento muito rápido ou um *faster pay back*. O propósito desta arquitetura é a construção de um DW incremental a partir de *Data Marts* independentes.

Comentários:

Acompanhando a figura da esquerda para a direita, é possível visualizar que os dados são extraídos de diversas fontes (OLTP, ORACLE, SQL SERVER, MYSQL, DB2, SAP, etc) e são carregados em um Data Warehouse. Em seguida, eles são extraídos, transformados e carregados em vários Data Marts. Logo, trata-se de uma abordagem Top-Down (DW > DMs)! A partir do DW são extraídos os dados e metadados para os Data Marts. Nos Data Marts as informações estão em maior nível de sumarização e, normalmente, não apresentam o nível histórico encontrado no DW. *Trata-se de abordagem de Inmon ou Kimball?* Trata-se da abordagem de Inmon!

Gabarito: Letra D

#VamosPraticar#

Referências e links deste capítulo

1 http://www.instagram.com/professordiegocarvalho

2 Não há consenso na literatura sobre os tipos de análise suportadas pelo BI, mas algumas questões não consideram que análises prescritivas são suportadas por essa tecnologia.

3 Granularidade é o nome que se dá ao nível de sumarização dos elementos de dados de um Data Warehouse. Quanto maior a granularidade, maior será o grão e menor será o nível de detalhes; quanto menor a granularidade, menor será o grão e maior será o nível de detalhes.

4 https://pt.wikipedia.org/wiki/Data_warehouse

5 https://pt.wikipedia.org/wiki/Ralph_Kimball